THE DETECTION OF SEX-LINKAGE IN FAMILIES COLLECTED AT RANDOM

CHAS. W. COTTERMAN
Genetics Laboratory, Ohio State University

For the genetic study of human characters two methods of collecting families are generally employed. If the character is quite rare the practice of studying only families containing at least one affected child is advisable. The small additional amount of information to be gained through the inclusion of families lacking affected members can in no way compensate for the greatly increased expenditure of time and effort. The method has a distinct advantage over that of random collection of families, since for the simpler types of hereditary behavior it will often be true that only one genotypic combination in the parents can produce affected children, in which cases the estimation of gene frequency and the necessary assumption of an equilibrium population are not required. Where the incidence of the trait is sufficiently high, on the other hand, a random collection of families in which the constitution of the families is disregarded will be found profitable. Moreover, for characters which are not readily recognized or generally reported, or those in which a special test is required, as in blood reactions and taste tests, a random collection cannot well be avoided.

Methods for the analysis of sex-linkage in families selected because at least one child is affected have been given by Hogben (1932), whereas an analysis applicable to randomly collected families has not yet been presented. An additional criterion for sex-linkage, however, has been suggested by Snyder (1934b), who shows that from a random survey of a population, separate estimates of the frequency of the sex-linked recessive gene may be gotten from the proportion of affected males and from the proportion of affected females, the test consisting in a comparison of the two statistics so obtained.

With respect to a sex-linked recessive character the expected proportions of affected sons and affected daughters among children of the various family types are easily deduced in terms of the frequency of the sex-linked recessive gene, q. Where these proportions are dependent upon q, the proportion expected in each case is found by application of the principle of random mating to be $q/1+q$. Designating this proportion by ζ for the

sake of simplicity, the offspring proportions for the four family types are as shown in Table I.

TABLE I

Type of Family	Expected Proportion of Affected Individuals	
	Among Sons	Among Daughters
i. Both parents normal	ζ	0
ii. Father only affected	ζ	ζ
iii. Mother only affected	1	0
iv. Both parents affected	1	1

The estimate of ζ is obtained from the observed frequencies of affected males and affected females in the general population from which the families are drawn. In a population at equilibrium with respect to the distribution of a sex-linked pair of alleles the proportions of normal and affected males are (1—q) and q, respectively, while for the females the proportions are ($1-q^2$) normal and q^2 affected females. Indicating the observed proportions in these classes by A, B, C, and D, respectively, and replacing q by $\zeta/1-\zeta$, we have the following equations of estimation:

$$A=1-2\zeta/1-\zeta,\ B=\zeta/1-\zeta,\ C=1-2\zeta(1-\zeta)^2,\text{ and } D=\zeta^2/(1-\zeta)^2 \ldots (1)$$

The probability of observing α normal and β affected individuals in a sample of N_1 males and of observing γ normal and δ affected individuals in a sample of N_2 females may therefore be written

$$P=\frac{N_1!N_2!}{\alpha!\beta!\gamma!\delta!}\left(\frac{1-2\zeta}{1-\zeta}\right)^{\alpha}\left(\frac{\zeta}{1-\zeta}\right)^{\beta}\left(\frac{1-2\zeta}{(1-\zeta)^2}\right)^{\gamma}\left(\frac{\zeta^2}{(1-\zeta)^2}\right)^{\delta}, \ldots\ldots (2)$$

or
$$L=\log P=\log N_1!\,N_2!/\alpha!\,\beta!\,\gamma!\,\delta!+(\alpha+\gamma)\log(1-2\zeta)+(\beta+2\delta)\log\zeta-(\alpha+\beta+2\gamma+2\delta)\log(1-\zeta)\ldots\ldots (3)$$

The optimum estimate of ζ from the observed N_1+N_2 individuals is obtained by employing Fisher's (1921) Method of Maximum Likelihood, which consists in finding the value of ζ for which L is maximum. Setting

$$\frac{dL}{d\zeta}=\frac{\alpha+\beta+2\gamma+2\delta}{1-\zeta}+\frac{\beta+2\delta}{\zeta}-\frac{2(\alpha+\gamma)}{1-2\zeta}=0\ldots\ldots\ldots\ldots\ldots (4)$$

leads to a quadratic equation of which the positive root is the maximum likelihood estimate,

$$(\zeta)_1 = \frac{\sqrt{(\alpha+2\beta+4\delta)^2+8\gamma(\beta+2\delta)} - (\alpha+2\beta+4\delta)}{4\gamma} \quad \ldots\ldots\ldots\ldots (5)$$

The notation (ζ) is used throughout this note to indicate a particular estimate of the parameter ζ. This function, (5), though somewhat laborious to calculate is nevertheless to be preferred over other estimates of ζ, since the invariance or reciprocal of the variance of this statistic, which is the amount of information respecting ζ elicited by the likelihood statistic in the observation of N_1+N_2 individuals, is never exceeded by that of any other estimate. This property of maximum likelihood statistics, which Fisher has termed "efficiency," is very readily demonstrated for this problem. The amount of information respecting any parameter, ζ, furnished by a sample of N_1+N_2 individuals is given by

$$I\zeta = S\left\{\frac{1}{m}\left(\frac{dm}{d\zeta}\right)^2\right\},$$

where m is the expected number in any class, the summation extending over all classes. The evaluation of this quantity for our problem may be presented in tabular form (Table II).

TABLE II

Expected Frequency, m	Differential coefficient, $dm/d\zeta$	$\frac{1}{m}\left(\frac{dm}{d\zeta}\right)^2$	
$N_1\frac{1-2\zeta}{1-\zeta}$	$-N_1\frac{1}{(1-\zeta)^2}$	$\frac{N_1}{(1-2\zeta)(1-\zeta)^3}$	$\frac{N_1}{\zeta(1-\zeta)^2(1-2\zeta)}$
$N_1\frac{\zeta}{1-\zeta}$	$N_1\frac{1}{(1-\zeta)^2}$	$\frac{N_1}{\zeta(1-\zeta)^3}$	
$N_2\frac{1-2\zeta}{(1-\zeta)^2}$	$-2N_2\frac{\zeta}{(1-\zeta)^3}$	$\frac{4N_2}{(1-2\zeta)(1-\zeta)^4}$	$\frac{4N_2}{(1-\zeta)^2(1-2\zeta)}$
$N_2\frac{\zeta^2}{(1-\zeta)^2}$	$2N_2\frac{\zeta}{(1-\zeta)^3}$	$\frac{4N_2}{(1-\zeta)^4}$	
N_1+N_2	0	$I\zeta = \frac{N_1+4N_2\zeta}{\zeta(1-\zeta)^2(1-2\zeta)}$	

This quantity is in fact the amount of information utilized by the likelihood estimate $(\zeta)_1$, the variance of which is therefore

$$V(\zeta)_1 = \frac{\zeta(1-\zeta)^2(1-2\zeta)}{N_1+4N_2\zeta} \quad \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

Putting N_1 and N_2 each equal to unity, the amount of information provided by an observation of one male and one female is

$$i(\zeta)_1 = \frac{1+4\zeta}{\zeta(1-\zeta)^2\,(1-2\zeta)}.$$

Now from the observed proportions, A, B, C, and D, it is possible by employing the relations in (1) to invent innumerable estimates of ζ, and the sampling variance of each statistic, (ζ), may be derived by employing the formula,

$$\frac{1}{N}V(\zeta) = S\left\{w\left(\frac{\partial(\zeta)}{\partial\nu}\right)^2\right\} - \left(\frac{\partial(\zeta)}{\partial N}\right)^2,$$

in which w is the expected proportion in any class, and ν the observed number in that class. Each of the frequencies ν is replaced by the expectancy Nw after differentiation and the expressions for all classes are summed. By putting N_1 and N_2 each equal to unity in the reciprocal of the variance formula thus obtained we have the amount of information $i(\zeta)$ respecting ζ elicited by the estimate through the observation of one male and one female. By dividing the amount of information elicited by the amount available, $i(\zeta)_1$, we have the proportional amount of information utilized by the alternative statistic. This proportion is termed the relative efficiency of the estimate in question and may be designated by $E(\zeta)$. Below are listed some of these other estimates together with the corresponding values of $E(\zeta)$.

In estimating ζ we might utilize either the male proportions (A and B) or the female proportions (C and D) separately. Thus

$$(\zeta)_2 = \frac{B}{1+B} = \frac{\beta}{\alpha+2\beta}, \text{ and } E\,(\zeta)_2 = \frac{1}{1+4\zeta}.$$

$$(\zeta)_3 = \frac{\sqrt{D}}{1+\sqrt{D}} = \frac{\sqrt{\delta}}{\sqrt{\gamma+\delta}+\sqrt{\delta}}, \text{ and } E\,(\zeta)_3 = \frac{4}{1+4\zeta}.$$

The estimates $(\zeta)_2$ and $(\zeta)_3$ are based on males alone and on females alone. The effect of putting N_1 and N_2 equal to unity

is to make $i(\zeta)_2$ and $i(\zeta)_3$ represent the amount of information per male and the amount per female, respectively. These values are compared with $i(\zeta)_1$, the information furnished by two individuals, one male and one female. $E(\zeta)_2$ and $E(\zeta)_3$ therefore give the proportional amounts of information used when equal numbers of both sexes are available but where females and males, respectively are disregarded in obtaining the estimate of ζ.

By averaging the two estimates thus derived we might obtain a fourth statistic,

$$(\zeta)_4=\frac{1}{2}\left(\frac{B}{1+B}+\frac{\sqrt{D}}{1+\sqrt{D}}\right)=\frac{\alpha\sqrt{\delta}+3\beta\sqrt{\delta}+\beta\sqrt{\gamma+\delta}}{2(\alpha+2\beta)\,(\sqrt{\gamma+\delta}+\sqrt{\delta})},$$

$$\text{and } E\,(\zeta)_4=\frac{16\zeta}{(1+4\zeta)^2}$$

Some other estimates and their efficiencies are

$$(\zeta)_5=\frac{D(1-B)}{B(1-D)}=\frac{\alpha\delta}{\beta\gamma}, \text{ and } E\,(\zeta)_5=\frac{\zeta(1-2\zeta)^2}{(1+4\zeta)\,(1-\zeta+\zeta^2)}.$$

$$(\zeta)_6=\frac{B-D}{1-D}=\frac{\beta\gamma-\alpha\delta}{\gamma(\alpha+\beta)}, \text{ and } E\,(\zeta)_6=\frac{(1-2\zeta)^2}{(1+\zeta)\,(1+4\zeta)}.$$

$$(\zeta)_7=\frac{D}{B+D}=\frac{\delta(\alpha+\beta)}{\beta(\gamma+\delta)+\delta(\alpha+\beta)}, \text{ and } E\,(\zeta)_7=\frac{1+\zeta}{1+4\zeta}.$$

The efficiency of each of the alternative estimates never exceeds unity and for many values of q the alternative estimates allow considerable loss of information. The values of the efficiencies of the alternative estimates in terms of per cent for values of q from 0 to 1 are shown in Figure 1.

SUMMARY

For a character attributable to a recessive autosomal gene substitution, the expected proportion of affected children among offspring of parents one of whom is affected is also $q/1+q$ or ζ, and the expected proportion of affected children among offspring of both normal parents is $(q/1+q)^2$ or ζ^2 (Snyder, 1934a). In terms of the gene-frequency, q, the proportion of affected individuals in the general population is the same as that for the female population with respect to a sex-linked recessive character and ζ is efficiently estimated as

$$(\zeta)=\sqrt{B'}/1+\sqrt{B'},$$

while

$$(\zeta^2)=B'/(1+\sqrt{B'})^2,$$

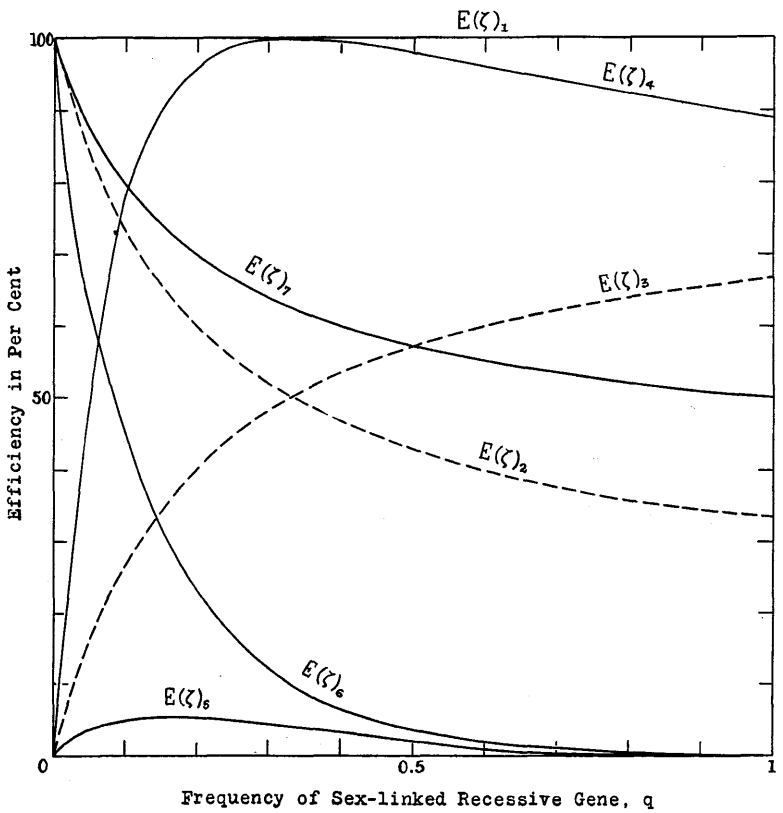

FIG. 1. The relative efficiencies of various estimates of the proportion ζ (=q/1+q) for a sex-linked recessive gene substitution. The top line, E $(\zeta)_1$, indicates that the likelihood estimate uses the whole of the available amount of information for all values of q.

where B' is the observed proportion of affected individuals in the general population. The variance of these proportions, however, may be written in somewhat simpler form than the formulae given by Snyder. As is shown in Table II, the variance of ζ is

$$V(\zeta) = \frac{(1-2\zeta)\,(1-\zeta)^2}{4N},$$

in which N is the total number of individuals in the sample. The variance of ζ^2 is readily derived by means of the transformation formula,

$$I\zeta^2 = \left(\frac{d\zeta}{d\zeta^2}\right)^2 I\zeta = \frac{1}{4\zeta^2} I\zeta.$$

Hence $$V(\zeta^2) = 4\zeta^2 \,.\, V(\zeta).$$

Employing these formulae the values of $V(\zeta)/N$ and $V(\zeta^2)/N$ for values of B' from 0 to 1 might readily be added to the tables of ζ and ζ^2 already given by Snyder, thus further simplifying calculations.

For a sex-linked recessive character, ζ is the proportion of affected individuals to be expected among children of normal mothers and affected fathers as well as the expected proportion of affected sons where parents are both normal. The problem of estimating ζ for a sex-linked character is, however, a somewhat more difficult task arithmetically, the optimum estimate and its sampling variance being obtained by means of equations (5) and (6) of this note.

REFERENCES

Fisher, R. A. 1921. On the mathematical foundations of theoretical statistics. Philos. Trans. Royal Soc. London, A, 222: 309–368.

1935. Statistical Methods for Research Workers. Edinburgh. Oliver and Boyd.

Hogben, Lancelot. 1932. The genetic analysis of familial traits III. Matings involving one parent exhibiting a trait determined by a single gene substitution with special reference to sex-linked conditions. Jour. Genet. 25: 293–314.

Snyder, L. H. 1934a. Studies in human inheritance X. A table to determine the proportion of recessives to be expected in various matings involving a unit character. Genetics 19: 1–17.

1934b. Modern analysis of human pedigrees. Eugen. News 19: 61–69.